Efficiency considerations for face recognition algorithms

NIST

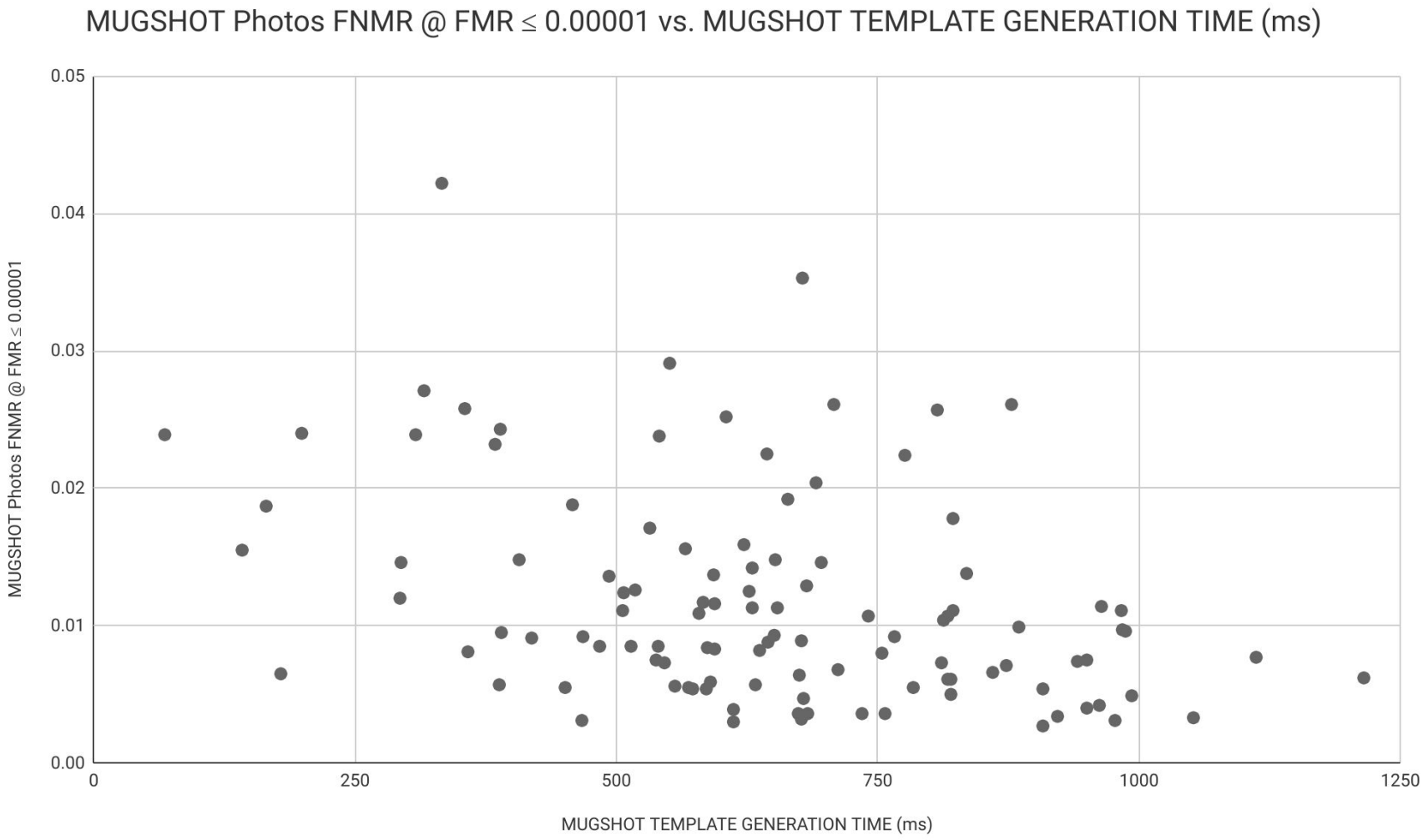
International Face Performance Conference (IFPC)
29 October, 2020

Presented by:
Brendan F. Klare, Ph.D.

Co-Founder & CEO
Rank One Computing

Denver, Colorado
brendan@rankone.io

Template Generation Speed vs. Accuracy



Little correlation between speed and accuracy

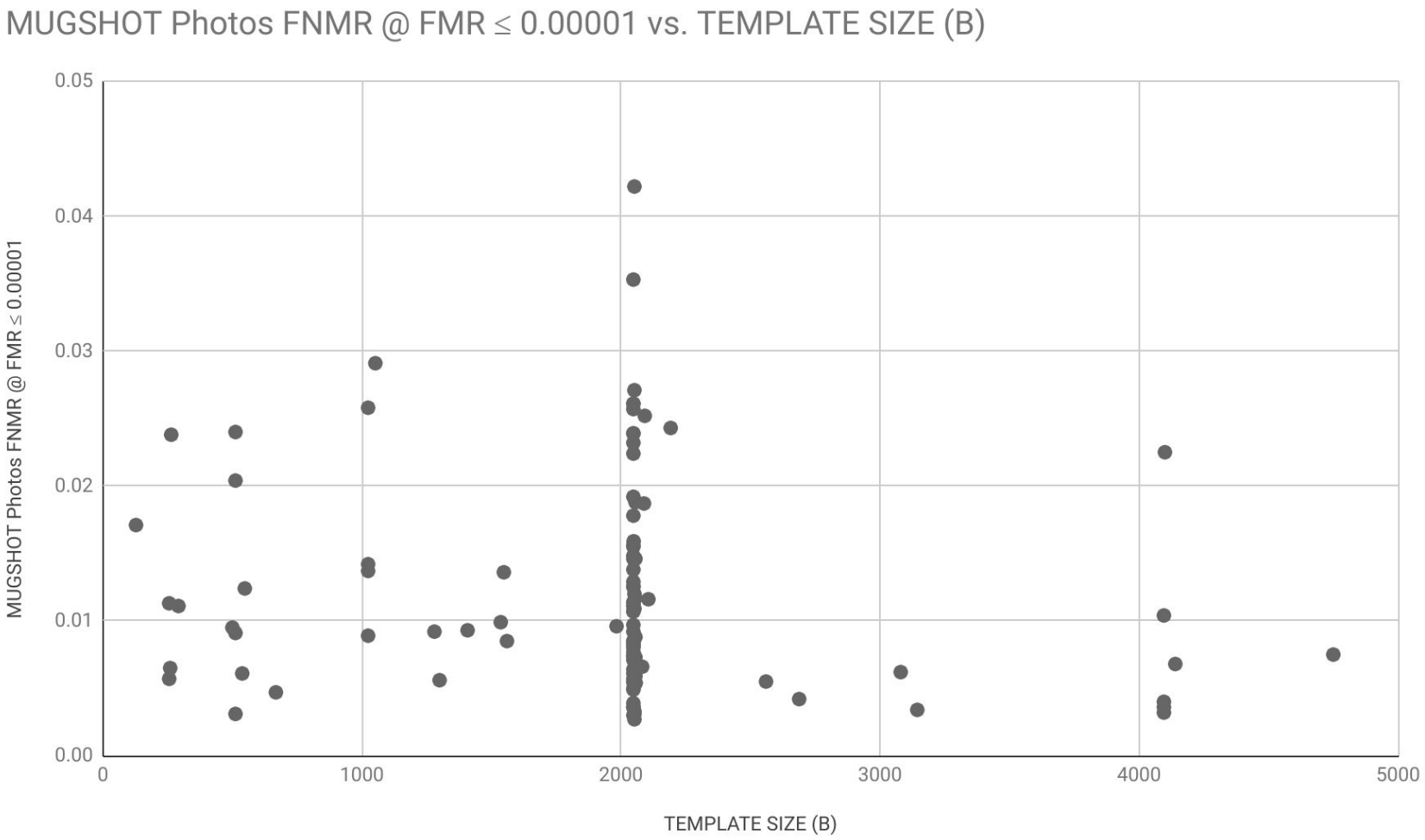
One generally does not need to sacrifice speed for accuracy

Over 5x difference in speed

Amongst high accuracy vendors (< 1% error rate), there is a massive difference in template generation speed

NIST FRVT Ongoing, July 27, 2020

Template Size vs. Accuracy



No correlation between template size and accuracy

No need to sacrifice template size for accuracy

Over 20x difference in template sizes

There is over an order of magnitude difference in template size for top-tier accuracy vendors

LET'S BREAK IT DOWN

Hardware Components

The primary limitations on Facial Recognition applications and their viability come from the host machine's processor and memory capacities.

The Processor (CPU)

Affects Speed

Template Generation	Comparison / Search
FAST 200ms **1 CPU** core can create **5 templates** per sec	**FAST** 1µs **1 CPU** core can perform **1,000,000 comparisons** per sec
SLOW 1s **1 CPU** core can create **1 templates** per sec	**SLOW** 10µs **1 CPU** core can perform **100,000 comparisons** per sec

The Memory (RAM)

Affects Efficiency

Database / Template Size	Model and Binaries
EFFICIENT 256 bytes **4GB of RAM** can host **~15,636,000 templates**	**EFFICIENT** 50 megabytes FR Software **loads immediately** & **doesn't affect system resources**.
INEFFICIENT 4096 bytes **4GB of RAM** can host **~976,000 templates**	**INEFFICIENT** 2 gigabytes FR software is **slow to load** & has a **noticeable impact on system resources**.

THE RIGHT HARDWARE FOR THE RIGHT APPLICATION

Architecture Options

Here is a description of this distribution

Embedded Device
Portability & Single-Purpose

PROCESSOR

1-8 cores
ARM, 1-2 GHz

RAM

1GB - 8GB

Desktop / Server
Dedicated Workstations & Workhorses

PROCESSOR

4 - 64 cores
x64 with 2 to 4 GHz

RAM

8GB - 256GB

Scalable Cloud
Flexible & Service Based

PROCESSOR

1 - 1000+ cores
x64 with 2 to 4 GHz

RAM

1GB - 16GB

PRIMARY USE CASE

Identity Verification

User Presents the system with:

Personal Identifier
ie. ID #, Keycard, Username

Face Capture
via Sensor (camera)

Retrieve Stored Template

Process Face
Detection > Templatization > Assess Quality > Assess Liveness

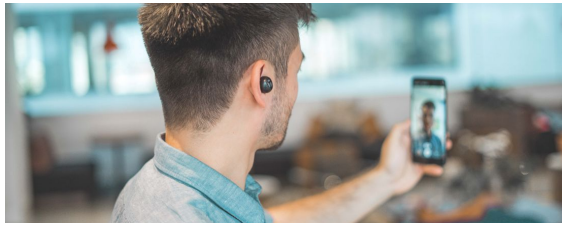
Compare Template Similarity

Access Granted

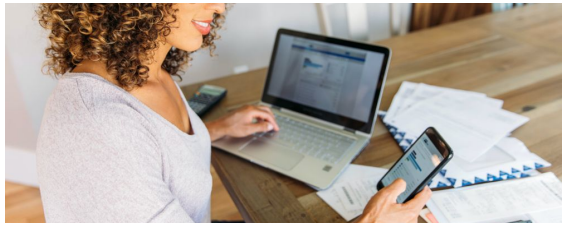
Access Denied

Bank Account Access

Secure Facility Access



Phone Unlock



Tax Return Filing

PRIMARY USE CASE

1:N Search

- Manually Performed by Trained Analyst / Examiner
- Automated by the Algorithm

Enter Probe Image

↓

Process Face
Detection > Templatization > Assess Quality

→

Database Search
From Identity Database

→

Identity Candidate List
with Similarity Scores

↓

Face Image Verification & Adjudication

→

Investigative Report

Identification of a Bank Robber
from a surveillance video frame.

Identification of an Assaulter
from their online dating profile.

Identification of a Hit & Run Suspect
from a bystander's cell phone camera

PRIMARY USE CASE

Automated Search

Person

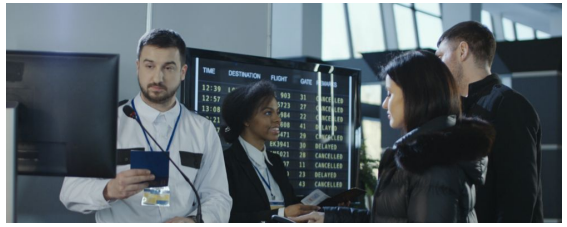
Captured by Sensor
Usually a Camera

Process Face
Detection > Templatization > Assess Quality > Assess Liveness

Database Search
From Identity Database

Similarity Threshold:

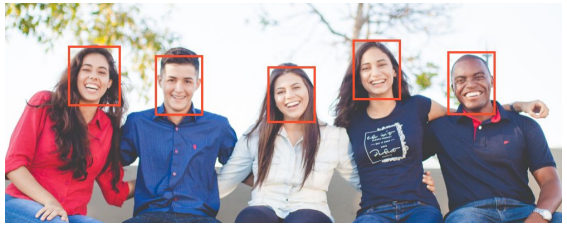
Match | **No Match**



Airport Security Screening

Home Security

Terrorist Watchlisting



Automated Photo Tagging

TAXONOMY OF BOTTLENECKS

Architecture vs. Use Case

	IDENTITY VERIFICATION **1:1**	MANUAL SEARCH **1:N**	AUTOMATED SEARCH **1:N+1**
Persistent Server/Desktop	• Slow template generation speed will reduce throughput/system response time. • Large binary size will impact system restart speed	• Large template size will require significant memory resources • High template generation speed will delay search results • High comparison speed will delay search results	• High template generation speed will reduce throughput (e.g., video processing) • Large template size will exasperate memory resources
Embedded Device	• Slow template generation speed will cause major latency (> 3 seconds). • Large binary size will occupy a high percentage of available memory.	• Template size must be very small due to memory limits • High template generation speed will significantly delay search results • High comparison speed will significantly delay search results	• High template generation speed will render video processing impossible • Template size must be very small due to memory limits
Scalable Cloud	• Large binary size will slow container instantiation time. • Poor network bandwidth will delay image transmission. • Slow template generation speed will reduce throughput / system response time.	• NOT ADVISED TYPICALLY • Large template size or large number of templates will make container instantiation very slow • Gallery size is typically too large to instantiate containers in less than 30 seconds	• Poor network bandwidth will prevent video transmission • High template generation speed make video processing expensive • Large template size,large number of templates, and/or large binary size will make container instantiation very slow

Identity Verification 1:1
Persistent Server / Desktop

- Slow template generation speed will reduce throughput/system response time
- Large binary size will impact system restart speed

Identity Verification 1:1
Embedded Device

- Slow template generation speed will cause major latency (> 3 seconds).
- Large binary size will occupy a high percentage of available memory

Identity Verification 1:1
Scalable Cloud

- Large binary size will slow container instantiation time
- Poor network bandwidth will delay image transmission
- Slow template generation speed will reduce throughput / system response time

Manual Search 1:N
Persistent Server / Desktop

- Large template size will require significant memory resources
- High template generation speed will delay search results
- High comparison speed will delay search results

Manual Search 1:N
Embedded Device

- Template size must be **very** small due to memory limits
- High template generation speed will significantly delay search results
- High comparison speed will significantly delay search results
- Large binary size will occupy a high percentage of available memory

Manual Search 1:N
Scalable Cloud

- **NOT ADVISED TYPICALLY**
- Large template size or large number of templates will make container instantiation very slow
- Gallery size is typically too large to instantiate containers in less than 30 seconds

Automated Search 1:N+1
Persistent Server / Desktop

- High template generation speed will reduce throughput (e.g., video processing)
- Large template and binary sizes will exasperate memory resources

Automated Search 1:N+1
Embedded Device

- Slow template generation speed will render video processing impossible
- Template size must be very small due to memory limits
- Large binary sizes will exasperate memory resources

Automated Search 1:N+1
Scalable Cloud

- Slow template generation speed make video processing expensive
- Large template size, large number of templates, and/or large binary size will make container instantiation very slow
- Poor network bandwidth will prevent video transmission

FRVT Wish List

- Require vendors to certify the algorithm they submit with their operationally available version
- Enable submission of an additional algorithm or two to help ensure operationally deployed algorithms can be NIST validated
    - E.g., some vendors deploy two different algorithms
- Benchmark speeds on ARM devices
- Be involved in crafting legislation / regulation
    - E.g., help require only NIST validated algorithms are used operationally
        - Amazon AWS and Clearview AI have never submitted their algorithms to FRVT
        - NEC and Microsoft have never submitted their algorithms to FRVT Ongoing

Presented by:
Brendan F. Klare, Ph.D.
Co-Founder & CEO
brendan@rankone.io

Rank One COMPUTING

BUILT BY ENGINEERS IN DENVER, COLORADO.

Read our blog for more information:
blog.rankone.io

Partnership Inquiries:
bd@rankone.io